Analyzing Privacy in Software

Feiyang Tang

Abstract

In our increasingly digital world, a pressing concern emerges: "How do we secure our privacy as we increasingly depend on software?" As we navigate through apps and platforms, the complexities of data privacy become evident. Understanding the intricate flow of personal data, ensuring compliance with evolving global regulations, and developing adaptable tools for diverse software environments are paramount. This Ph.D. thesis delves deep into these challenges, offering insights and solutions that span from the granular details of code to the broader validation of privacy policies.

The first challenge is the subtlety of personal data. Legal definitions are often abstract and translating them into technical requirements is no easy task. Identifying what constitutes personal data in a sea of code is a daunting challenge. Secondly, understanding how personal data flows within systems is crucial. With regulations like the General Data Protection Regulation (GDPR) in place, it is crucial to know what kind of processing personal data undergoes for compliance checks. Lastly, different projects have different needs. For developers doing self-analysis, a detailed examination of compiled code can reveal intricate data flows. However, for large industry projects, high-level source code analysis may be more practical for third parties to quickly gauge privacy compliance situations across millions of lines.

Investigations into these aspects resulted in the seven papers that are presented in this dissertation. They also led to the following additional contributions: (1) A privacy-flow-graph tailored for Java and Android applications; this approach aids in the Data Protection Impact Assessment (DPIA) process. (2) A biometric data identification approach developed to pinpoint biometric API usage within Java and Android applications; this method ensures alignment with the GDPR. (3) An automatic comparison approach that addresses the collection of user interaction data in mobile apps by comparing an app's privacy policy claims with its actual code implementation. (4) An automated code review assistant that offers a method to identify and categorize relevant code segments in source code, thus reducing the manual review effort.

The contributions offer guidance for developers and legal experts, connecting the detailed aspects of software development with the clear rules of privacy regulations. These contributions can pave the way for a clearer, more streamlined, and compliant online environment, ensuring that as we use digital platforms, our privacy is always protected.

List of papers

Paper 1

Tang, F. and Østvold, B. (2022). Assessing software privacy using the privacy flow-graph. In Proceedings of *the 1st International Workshop on Mining Software Repositories Applications for Privacy and Security (MSR4P&S 2022)*. Association for Computing Machinery, New York, NY, USA, 7–15.

Read the preprint here.

Paper 2

Tang, F. (2022)., PABAU: Privacy Analysis of Biometric API Usage, In Proceedings of *the 2022 IEEE Conference on Privacy Computing (PriComp 2022)*, Haikou, China, 2022, pp. 2301-2308.

Read the preprint here.

Paper 3

Tang, F.; Østvold, B. and Bruntink, M. (2023). Identifying Personal Data Processing for Code Review. In Proceedings of *the 9th International Conference on Information Systems Security and Privacy - ICISSP*; ISBN 978-989-758-624-8; ISSN 2184-4356, SciTePress, pages 568-575.

Read the preprint here.

Paper 4

Tang, F.; Østvold, B. and Bruntink, M. (2023). Helping Code Reviewer Prioritize: Pinpointing Personal Data and its Processing. In Proceedings of *the 22nd International Conference on Intelligent Software Methodologies, Tools and Techniques (SOMET 2023). DOI: 10.3233/FAIA230228.*

Read the preprint here.

Paper 5

Tang, F. and Østvold, B. (2023). Transparency in App Analytics: Analyzing the Collection of User Interaction Data. In Proceedings of *the 20th Annual International Conference on Privacy, Security & Trust (PST 2023).*

Read the preprint here.

Paper 6

Tang, F. and Østvold, B. (2023). User Interaction Data in Apps: Comparing Policy Claims to Implementations. Presented and submitted to *the 18th IFIP Summer School on Privacy and Identity Management 2023 (IFIPSC 2023).*

Paper 7

Tang, F. and Østvold, B. (2023). Helping Privacy Reviewers Identify and Categorize Relevant Code. To be submitted to *the IEEE International Conference on Software Analysis, Evolution and Re-engineering (SANER 2024).*

Paper 8

Tang, F. and Østvold, B. (2023). Software Privacy and Program Analysis: Insights, Methods, and Opportunities. A book chapter intended for submission to the Springer Handbook on *Privacy and Security Matters in Biometric Technologies*.